# Der komplette Guide zur Datenintegration

## Einfache Datenintegration im digitalen Zeitalter

E-BOOK

data virtuality

any data.
any system.
any time.

# Der komplette Guide zur Datenintegration

*Einfache Datenintegration im digitalen Zeitalter*

**ZIELGRUPPE:**

Business Intelligence-Manager
Data Warehouse-Spezialisten
Projektmanager
Solution Architects

## Die Herausforderung: Das Potenzial von Big Data ausschöpfen

**Big Data ist mittlerweile Realität – und verändert die Business-Welt von Grund auf. Big Data liefert neue Einblicke und beschleunigt die Generierung von Business-Informationen. Das Konzept ist nicht neu, doch sein Potenzial für die effiziente Datenorganisation, -verwaltung und -analyse wird gerade erst entdeckt. Riesige Mengen an intern und extern generierten Daten stehen endlich für die allgemeine Nutzung zur Verfügung.**

Dennoch: Für viele Unternehmen ist es nicht einfach das Potenzial dieser Daten auszuschöpfen. Auch wenn der Nutzen von Big Data zum Teil gerade im Volumen und in der Geschwindigkeit liegt, mit der diese Daten generiert werden – genau diese Merkmale verursachen die meisten Schwierigkeiten. Noch entmutigender ist die große Bandbreite an Dateitypen und Dateiquellen, die von stark strukturierten Dateien über halbstrukturierte Texte bis hin zu unstrukturierten Video- und Audio-Feeds reicht.

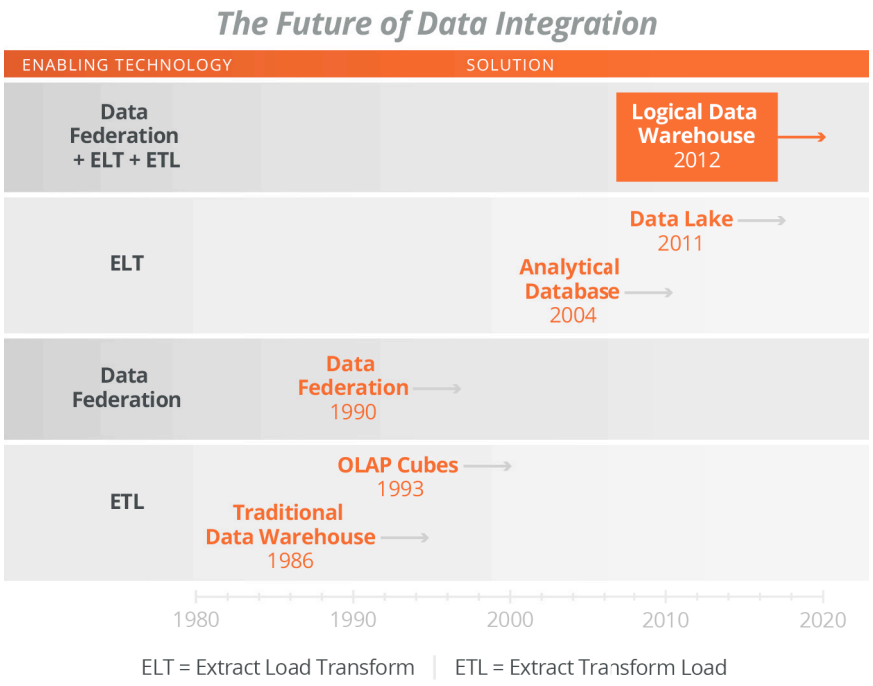
## *Die größten Herausforderungen von Big Data*


In einer durchgeführten Studie des IT-Analysten Gartner berichteten 49 % der Organisationen, dass sie am meisten mit der Vielfalt von Big Data zu kämpfen hätten. 35 % bezeichneten das Datenvolumen als ihr größtes Problem. 16 % der Unternehmen erklärten, dass ihnen das Tempo, mit der die Daten generiert würden, die meisten Schwierigkeiten bereite[1]. Das Handling von Daten aus verschiedenen Datenbanken und Systemen war schon immer eine Herausforderung – mit der wachsenden Zahl unterschiedlicher Dateitypen lässt sich dieses Problem aber kaum mehr bewältigen.

Hinzu kommt, dass die Daten oft auf mehrere separate Systeme, Quellen und Silos verteilt sind. Da kann eine unternehmensweit einheitliche Sicht auf die verfügbaren Informationen als reine Utopie erscheinen.

Für Unternehmen, die diese Datenflut auf die gleiche Weise wie vor 20 Jahren integrieren möchten – nämlich mit einem traditionellen Data Warehouse-Konzept – ist es tatsächlich (fast) ein Ding der Unmöglichkeit. Um wirklich einen Nutzen aus diesen Daten ziehen zu können, müssen Informationen aus internen UND externen Quellen aufgenommen und verarbeitet sowie echtzeitnahe Analysen durchgeführt werden. Kein leichtes Unterfangen. Angesichts dieser Herausforderungen können traditionelle Data Warehouse-Lösungen nicht mit der rasanten Entwicklung der Daten-Ökosysteme Schritt halten.

1 Gartner 2014: “Survey Analysis: Big Data Adoption in 2013”




# Das traditionelle Data Warehouse

In einer typischen IT-Umgebung übernehmen, modellieren und speichern traditionelle Data Warehouses die Daten im Rahmen eines Extraktions-, Transformations- und Ladeprozesses (ETL). Solche ETL-Jobs verschieben große Datenmengen in einer Batch-orientierten Arbeitsweise und das in der Regel jeden Tag. Die tägliche Ausführung dieser Prozesse bedeutet im besten Falle, dass die Warehouse-Daten einige Stunden, meistens aber einen Tag oder älter sind. Eine häufigere Ausführung lässt sich nur schwer rechtfertigen, da die ETL-Jobs erhebliche CPU-, Arbeitsspeicher-, Festplatten- und Netzwerkkapazitäten beanspruchen. Als APIs noch nicht so flächendeckend eingesetzt wurden wie heute, waren ETL-Tools die Lösung der Wahl für den operativen Einsatz. Jetzt, da es APIs und damit eine riesige Vielfalt an Daten gibt, ist die ETL-Methode nicht mehr praktikabel.

Doch schon in den Zeiten vor API und Big Data stellten ETL-Tools eine große Herausforderung dar, denn sie verlangten umfassende Kenntnisse über jede operationale Datenbank oder Anwendung.

Interkonnektivität ist sehr komplex und erfordert ein fundiertes Wissen zu jeder einzelnen Datenquelle ¬– bis hinunter zur Feldebene. Je größer die Zahl der miteinander verbundenen Systeme, die in das Data Warehouse integriert werden sollen, desto komplizierter ist diese Aufgabe.

***Schneller als je zuvor tauchen im digitalen Zeitalter neue Anforderungen auf – und die alten ändern sich. Deshalb sind Agilität und Reaktivität erfolgsentscheidende Faktoren.***

So kommt es, dass Data Warehouse-Projekte mittlerweile im Ruf stehen, eine erschreckend hohe Misserfolgsquote zu haben. Wenn sie nicht direkt scheitern, verursachen sie häufig Mehrkosten und Verzögerungen bei der Implementierung. Große Sorgfalt ist bei der Datenbankkonzeption und der Definition der Anforderungen gefragt, um komplizierte und fragile Verbindungen nicht neu bearbeiten zu müssen. Denn selbst kleinste Änderungen haben aufgrund enger Interdependenzen oft unvorhersehbare und weitreichende Konsequenzen.

Ein weiterer Nachteil des ETL-Warehouse-Konzepts: Die Mitarbeiter bekommen die Ergebnisse selten vor Abschluss des mehrmonatigen Entwicklungsprozesses zu sehen. Bis zu diesem Zeitpunkt haben sich die Anforderungen aber oft schon geändert. Manchmal sind auch Fehler entdeckt worden oder die Projektziele haben sich verlagert. Jeder einzelne dieser Faktoren kann die IT-Abteilung zurück an den Schreibtisch zwingen, um die neuen Anforderungen zusammenzustellen – und aller

Wahrscheinlichkeit nach mehrere Monate Arbeit über Bord zu werfen. Laut Schätzungen von Gartner liefern zwischen 70 und 80 Prozent der durchgeführten Business Intelligence-Projekte nicht die erwarteten Ergebnisse[2].

Ursprünglich wurden Data Warehouses eher für das betriebliche Reporting statt für die interaktive Datenanalyse entwickelt. Um ein traditionelles Data Warehouse für Analyseabfragen zu nutzen, muss man sehr sorgfältig eine passende Struktur entwickeln und umfassende, spezifische Performance-Optimierungsmaßnahmen durchführen. Wer die Daten später anders nutzen möchte, muss die Datenstruktur ändern und neu optimieren – ein umständlicher und kostenaufwendiger Prozess.

Das grundsätzliche Problem des traditionellen ETL-Ansatzes liegt jedoch in der schieren Anzahl verfügbarer Datenquellen und den unzähligen Möglichkeiten des Datenzugriffs – z. B. in der Zunahme von APIs. APIs basieren auf Datenimport und -export und besitzen immer ein eigenes Zugriffsprotokoll. Technisch ist es zwar möglich, diese Art von Konnektivität mit ETL zu implementieren, doch die Implementierung ist extrem komplex, schwer zu managen und teuer zu erweitern. Und diese Probleme sind noch gravierender, wenn die APIs keine Datenaustauschstandards wie ODBC oder JDBC verwenden.

Schneller als je zuvor tauchen im digitalen Zeitalter neue Anforderungen auf – und die alten ändern sich. Deshalb sind Agilität und Reaktivität erfolgsentscheidende Faktoren. Traditionelle Data Warehouses können mit den Anforderungen moderner Unternehmen und den Trends des digitalen Wandels schlicht und ergreifend nicht Schritt halten. Diese Defizite waren der Grund für die Entstehung neuer Datenverarbeitungsmethoden, und der nächste Ansatz, der entwickelt wurde, hieß Multidimensionales Online Analytical Processing (OLAP).

2 Quelle: http://www.computerweekly.com/news/1280094776/Poor-communication-to-blame-for-business-intelligence-failure-says-Gartner

### DAS TRADITIONELLE WAREHOUSE AUF EINEN BLICK

| + | – |
|---|---|
| • Bewegen großer Datenmengen<br>• Konzipiert für das betriebliche Reporting | • Erhebliche Belastung von Netzwerk, CPU usw.<br>• Lange Entwicklungszyklen (mehrere Monate)<br>• Keine interaktive Datenanalyse<br>• Hochkomplexe Implementierung durch viele verschiedene Daten-integrationsmöglichkeiten |

# Multidimensionale Datenbanken/ Modellierung (Cubes)

## OLAP

Online Analytical Processing (OLAP) und Cubes stehen für multidimensionale Datenbestände, die im Wesentlichen als Zwischenspeicherbereich für die Analyse dienen. Diese speziellen OLAP-Datenbanken speichern Daten nicht in Tabellen, sondern in OLAP-Cubes – eine Methode der Datenspeicherung und -abfrage anhand einer organisierten, multidimensionalen Struktur, die eigens für die Analyse optimiert wurde.

OLAP-Datenbanken sind so konzipiert, dass sie möglichst viele Abfragen und Kombinationen von Datenfeldern im Voraus berechnen. Auf diese Weise liefern sie schnelle Abfrageergebnisse. Doch auch wenn diese Lösungen effizienter arbeiten als klassische relationale Datenbanken: Ihre multidimensionale Struktur macht sie unflexibel, und

Änderungen können nicht ohne weiteres integriert werden. Hinzu kommt, dass die Speicherung großer Datenmengen in einem Cube Performance-Engpässe produziert. Für einfache Anwendungsszenarien sind OLAP-Datenbanken durchaus sinnvoll, doch für große Datenbestände müssen parallel weitere Tools herangezogen werden – was die Analyse verkompliziert und ein spezifisches Know-how erfordert.

## ROLAP

Eine weitere Methode zur Organisation von Daten für die multidimensionale Abfrage ist Relational Online Analytical Processing (ROLAP). ROLAP ist eine Form des OLAP, mit der multidimensionale Analysen von Daten durchgeführt werden, die in einer relationalen statt in einer multidimensionalen Datenbank (die als OLAP-Standard gilt) gespeichert sind.

ROLAP ist bei der Verarbeitung großer Datenvolumina zwar leistungsfähiger als OLAP-Datenbanken, kann bei kleineren Datenmengen aber nicht mit der Geschwindigkeit und Effizienz eines OLAP-Systems mithalten. ROLAP-Datenbanken bringen einen hohen manuellen Wartungsaufwand mit sich und sind für Business-User schwer zu bedienen. Deshalb haben sie den Ruf, unflexibler zu sein als OLAP-Cubes. Sowohl OLAP- als auch ROLAP-Datenbanken sind derzeit noch weit verbreitet – aber keine der beiden Technologien erfüllt die heutigen Ansprüche im Hinblick auf echtzeitnahe Informationen für die Analyse und unstrukturierte Daten.

### MULTIDIMENSIONALE DATENBANKEN (OLAP, ROLAP) AUF EINEN BLICK

| + | – |
|---|---|
| • Speicherung und Abfrage in strukturierter Form<br>• Schnelle Abfrage-ergebnisse dank vorausberechneter Datenfelder<br>• Schnell und effizient bei kleinen Datenmengen | • Probleme bei großen Datenmengen<br>• Mangelnde Flexibilität durch multidimensionale Struktur<br>• Performance-Engpässe aufgrund der Speicher-limitierung von Cubes<br>• Manuelle Wartung<br>• Schwer zu bedienen für Business-User<br>• Zusätzliche Tools bei großen Datenvolumina notwendig |

## Self-Service Business Intelligence

Weder ein Data Warehouse noch die OLAP-Technologie konnten also die Erwartungen im Hinblick auf einen schnellen und umfassenden analytischen Datenzugriff erfüllen. Deshalb wurde ein neuer Ansatz entwickelt: SSBI-Technologien (Self-Service Business Intelligence) wie Qlik und Tableau brachten eine Methode der Datenanalyse hervor, mit der Business-User ohne Unterstützung ihrer IT-Kollegen auf Unternehmensdaten zugreifen und nutzen können. SSBI-Tools haben die Fähigkeit, Informationen aus Data Warehouses und Daten, die nicht im Data Warehouse gespeichert sind, zu „joinen“ bzw. lokal zu integrieren.

Dies geschieht, indem Kopien der Datenquellen in einen lokalen Datenspeicher gelegt werden, in dem der Analyst die Daten je nach Bedarf „joinen“ oder integrieren kann.

Diese Self-Service-Tools sind flexibel und relativ einfach zu implementieren – und machen den Datenanalysten weitgehend unabhängig. Doch sie haben auch eindeutig ihre Schwächen: Ihr Hauptnachteil liegt darin, dass so durchgeführte Datenanalysen in kurzer Zeit unüberschaubar werden. Auf breiter Ebene angewendet, verursachen sie die doppelte Arbeit, liefern uneinheitliche Ergebnissen und führen zu einer kurzen, chaotischen Berichterstattung. Jeder Nutzer kann eigene Regeln und Berechnungen erstellen; da kann es durchaus passieren, dass verschiedene Gruppen oder Mitarbeiter dieselben KPIs und Metriken auf unterschiedliche Weise berechnen. Das Resultat: unterschiedliche, miteinander kollidierende Ergebnisse und widersprüchliche Informationen.

Da diese Lösungen keine Zugriffsregelung aufweisen, gibt es zudem keine Sicherheitsebene für den Schutz sensibler Daten – eine echte Schwachstelle, denn Analysten tauschen häufig unbedacht Dateien aus. Auch die Möglichkeiten zur Datenumwandlung sind relativ begrenzt. Zudem führen viele Rechner parallel die gleichen Aufgaben für unterschiedliche Nutzer aus, so dass leistungsstarke Computerressourcen ineffizient genutzt werden. Das trägt zu höheren Kosten und einer niedrigeren Performance bei. Aus all diesen Gründen können reine SSBI-Tools zwar einen begrenzten, kurzfristigen Bedarf erfüllen, eignen sich aber nicht als End-to-End-Analyselösung auf Unternehmensebene.

### SELF-SERVICE-BI-TOOLS AUF EINEN BLICK

| + | – |
|---|---|
| • Business-User können ohne IT-Unterstützung Analysen durchführen<br>• Daten aus externen Quellen können mit Daten aus dem Data Warehouse „gemischt“ werden<br>• Flexibel und leicht zu implementieren | • Unterschiedliche KPI-Berechnungen wegen dezentraler Analysen<br>• Keine Sicherheitsebene<br>• Begrenzte Datentransformation<br>• Ineffiziente Ressourcen-auslastung wegen paralleler Nutzung |

## Analytische Datenbanken

Die SSBI-Tools entwickelten sich weiter, doch die Herausforderung blieb die gleiche: eine analytische Datenbank zu finden, die für die Analyse dieselbe Flexibilität bot wie relationale Datenbanken für die transaktionale Datenverarbeitung.

Progressive Softwarehersteller versuchten, die Grenzen der Data Warehouses, Cubes und SSBIs zu überwinden. Sie begannen, auf Datenbanken hinzuarbeiten, die nicht nur flexibel, sondern auch in der Lage waren, Analysedaten zu verarbeiten. Diese analytischen oder spaltenorientierten Datenbanken stellten den nächsten Schritt auf dem Weg zu einer zufriedenstellenden Lösung für Business-Analysten dar. Sie hatten sich zu analytischen MPP-Datenbanken (MPP = Massively Parallel Processing) entwickelt, die flexibler und leistungsfähiger als Cubes waren – sogar dann, wenn große Datenmengen gespeichert und abgefragt werden.

Bei analytischen Datenbanken müssen die Daten jedoch mit Prozessen einkopiert werden, die den oben beschriebenen ETL-Prozessen sehr ähnlich sind – und leider auch ähnliche Nachteile bergen. Die Ladeprozesse sind in der Regel langsamer als in einem traditionellen, datensatzbasiert arbeitenden Data Warehouse. Denn um die Daten für eine schnelle Analyseabfrage zu optimieren, ist ein zusätzlicher Schritt erforderlich.

Dieser Extraschritt ist notwendig, um die Daten von einem datensatzbasierten Format in ein Spaltenformat zu konvertieren und die Daten dann auf Feldebene zu komprimieren. Er bringt zwar erhebliche Performance-Verbesserungen mit sich, benötigt trotzdem mehr Zeit – Zeit, die dem Analysten bei der Datenanalyse fehlt. Diese Latenzzeit verhindert, dass aus analytischen Datenbanken Echtzeitdaten abgerufen werden können.

### ANALYTISCHE DATENBANKEN AUF EINEN BLICK

| + | – |
|---|---|
| • Skalierbarkeit und Verarbeitung großer Datenmengen<br>• Leistungsfähige parallele Verarbeitung<br>• Hohe Skalierbarkeit | • Langsame Ladeprozesse wegen Konvertierung von datensatz- in spalten-basierte Daten und Datenkomprimierung<br>• Kein Echtzeitzugriff<br>• Kein Agilität |

## Data Lakes und ELT

Als Nächstes wurde die Data-Lake-Strategie entwickelt. Data Lakes sind Datenbanken, in denen riesige Mengen von Rohdaten solange in ihrem nativen Format gehalten werden, bis man sie benötigt. In puncto Leistungsfähigkeit und Flexibilität stellen

Data Lakes – in vielen Fällen Hadoop-basierte Systeme – die nächste Entwicklungsstufe dar. Ein großer Vorteil dieser Systeme liegt darin, dass die Daten vor der Abfrage nicht strukturiert (transformiert), also nicht nach dem „Schema-on-write"-Prinzip abgespeichert werden müssen. Im Gegenteil, erst bei der Abfrage wird ihnen eine Struktur zugewiesen („Schema-on-Read"-Prinzip). In solchen „Datenseen" lassen sich tatsächlich große Mengen unstrukturierter Daten kosteneffizient halten. Doch für interaktive Analysen, bei denen schnelle Ergebnisse oder Echtzeitdaten benötigt werden, sind Data Lakes unzureichend.

Der vermehrte Einsatz von Data Lakes ermöglicht den Wechsel von ETL zu ELT (Extrahieren, Laden und Transformieren). Anders als ETL-Prozesse, bei denen die Daten vor dem Laden in die Datenbank umgewandelt werden, nehmen ELT-Prozesse die Daten im Rohzustand auf und verkürzen damit die Ladezeit erheblich. Der Gedanke dahinter: Data Lake-Speichertechnologien sind hinsichtlich der Datenstruktur nicht wählerisch. Deshalb ist kein Entwicklungsaufwand nötig, um die Daten vor der Abfrage und Analyse in die richtige Struktur zu überführen. Alle Daten werden einfach in den Data Lake „geschüttet" bzw. dort „geparkt". Jeder weitere Vorgang, jede Umwandlung kann innerhalb dieser Datenbank erfolgen – und zwar zu dem Zeitpunkt, an dem es erforderlich ist.

Data Lakes sind ein verlockendes Konzept. Doch leider halten sie nicht das, was sie versprechen, und zwar aus verschiedenen Gründen: Oberstes Ziel eines „Datensees" ist die Vereinfachung und Beschleunigung der Datenbankvorgänge. Doch häufig verkompliziert er sie durch Extra-Arbeitsschritte, mit denen die Daten für die Analyse aufbereitet werden. Und obwohl Data Lakes die Ladeprozesse erheblich erleichtern, müssen alle Daten nach wie vor an einen zentralen Ort verschoben oder kopiert werden, bevor sie zu Analysezwecken abgefragt werden können. Diesen Nachteil haben sie mit traditionellen Data Warehouses, die mit ETL arbeiten, gemeinsam. Denn die Latenzzeiten beim

Laden der Daten lassen sich nicht aus der Analysedaten-Lieferkette eliminieren, auch wenn diese im Vergleich zum Data Warehouse erheblich kürzer sind. Ein weiteres Manko des Data Lake sind die „Data Swamps“ (Datensumpf), auch „Data Graveyards“ (Datenfriedhof) genannt: Die Data Lake-Methode hat häufig den Effekt, dass wegen der niedrigeren Speicherkosten sehr viel mehr Daten in der Datenbank abgelegt werden als beim ETL-Konzept. Die Unternehmen laden und speichern erheblich mehr Informationen, als sie tatsächlich analysieren können. Jedes Laden nimmt aber Zeit, Speicher und Netzwerkressourcen in Anspruch. Somit verursachen unnötige Ladevorgänge Kosten und weitere Latenzzeiten – und verzögern die zeitnahe Verarbeitung analytisch wertvollerer Daten.

Data Lakes und ELT-Prozesse führen zwar die Daten an einem Ort zusammen – doch sie bieten weder schnelle Abfrageergebnisse wie analytische Datenbanken noch einen Echtzeitzugriff auf die Daten.

## DATA LAKES UND ELT AUF EINEN BLICK

| + | – |
|---|---|
| • Aufnahmen riesiger Mengen unstrukturierter Daten<br>• Daten müssen vor der Abfrage nicht strukturiert werden<br>• Effiziente Ladevorgänge | • Kein Echtzeitanalyse möglich<br>• Daten müssen vor der Analyse an einen zentralen Speicherort verschoben werden<br>• Niedrige Kosten begünstigen „Daten-friedhöfe“, die die Performance verringern und Kosten steigern |

# Moderne Datenintegrations-architekturen

Beim Rückblick auf das traditionelle Data Warehouse und die Data Lakes wird eine Gemeinsamkeit deutlich: Die Informationen befinden sich alle in einer zentralen, physischen Datenbank. Man glaubte, die Daten an einem zentralen Ort „einpferchen" zu müssen, bevor man mit ihnen arbeiten könne. Diese Annahme bremste die Beschleunigung des Datenzugriffs – und war der Pferdefuß bei allen anderen Methoden, die zuvor erläutert wurden.

# Data Federation – endlich eine Erleichterung

Während die Mehrheit der Datenanalysten damit beschäftigt war, die relationalen Datenbanken zu Cubes, analytischen Datenbanken und Data Lakes weiterzuentwickeln, versuchte ein anderes Lager, Daten mithilfe der Data Federation-Methode zu integrieren.

Data Federation ist eine Methode, mit der man Sofortabfragen mehrerer, voneinander getrennter Datenbanken durchführen kann, ohne die Daten aus den ursprünglichen operationalen Quellen in eine zentrale Analysedatenbank zu kopieren oder zu verschieben. Angesichts der Promptheit, mit der hier Daten analysiert werden konnten, stellte dieser Ansatz eine signifikante Verbesserung im Vergleich zu allen anderen Vorgängertechnologien dar.

Die Idee ist gut, ihr Nutzen unbestritten – doch Data Federation allein ist bei großen Datenmengen oder vielen parallel arbeitenden Nutzern keine skalierbare Lösung. Zudem ist sie stark abhängig von der Geschwindigkeit und Stabilität der Quellsysteme und des Netzwerks. Ihre Performance leidet in der Regel sowohl unter den Datenanalyse- als auch unter den

Produktionsprozessen. Data Federation ist also schnell und flexibel, doch an sich nicht skalierbar oder besonders verlässlich. Aber sie ist ein wichtiger Schritt in die richtige Richtung.

Die nächste Phase der Entwicklung bestand in einer Kombination der Data Federation-Methode mit Cache-Datenbanken, um die beschriebenen Probleme zu beheben. Dieses „hybride" Konzept griff auf Big-Data-Lösungen zurück, um das Data Warehousing zu ergänzen. Das Ergebnis war ein Mix aus Datenbanken, Virtualisierung und verteilten Prozessen für das Datenmanagement, das die größten Vorzüge verschiedener Technologien miteinander verband. Doch ein robustes, agiles und leistungsstarkes Data Warehouse konnte auch dieses Konzept nicht bieten. Caching kann problematisch sein, denn die Cache-Auslastung muss um die Performance der Quellsysteme herum geplant werden. Außerdem wird der Cache in eine Datenbank geladen, das nicht unbedingt für die Aufnahme verschiedener Datenbestände bzw. Datentypen optimiert wurde.

Um sich dem modernen Data Warehouse anzunähern, ist die virtuelle Datentechnologie von wesentlicher Bedeutung – von der einfachen Data Federation hin zur Virtualisierung, virtuellen Ansichten, Indizes und semantisch vereinheitlichten Datenzugriffsschichten. Die Entwicklung virtueller oder logischer Datenansichten nimmt weniger Zeit in Anspruch als die physische Verschiebung aller Daten und kann mühelos per Mausklick erfolgen. Obendrein können virtuelle Ansichten verändert werden, ohne – wie bei früheren Data Warehouse-Integrationsmethoden – die Daten umzuwandeln und neu zu laden. Das bedeutet: Die Änderungen lassen sich umgehend live darstellen, ohne eine Nacht lang auf das Laden der Daten zu warten. Die Virtualisierung der Datenintegration ermöglicht eine extreme Agilität in der Entwicklung und reduziert Build-Zeiten und -Kosten erheblich. Und sie führte zum nächsten Durchbruch im Data Warehousing.

### DATA FEDERATION AUF EINEN BLICK

| + | – |
|---|---|
| • Zusammenfassung von Datenbanken in einem zentralen Repository ohne Kopieren der Daten<br>• Sehr schneller Datenzugriff<br>• Flexible Änderung virtueller Ansichten<br>• Extreme Agilität und verringerte Build-Zeiten/-Kosten durch virtuelle Datenintegration | • Begrenzte Skalierbarkeit (z.B. viele simultane Nutzer)<br>• Caching-Datenbanken verursachen Performance-Probleme |

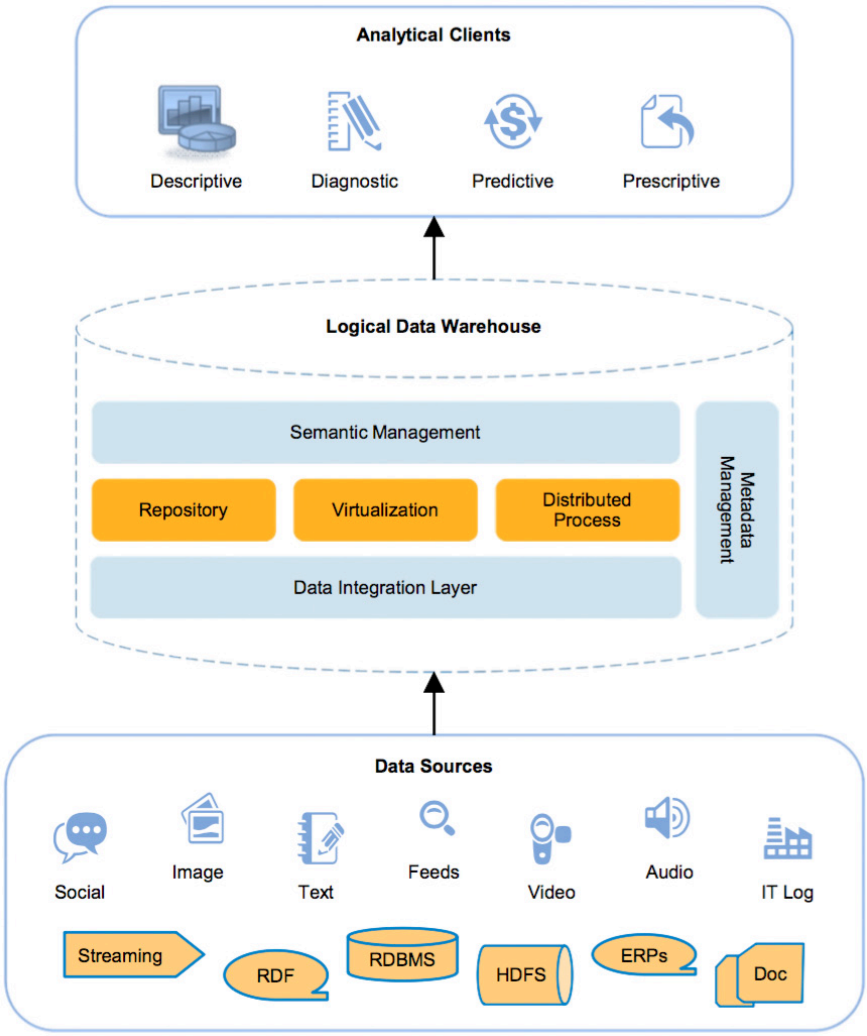
## DataVirtuality: Das erste Logical Data Warehouse

Moderne Datenintegrationsstrategien arbeiten nach dem „Best-Fit-Engineering“-Prinzip, d. h. jeder Teil der Datenmanagementinfrastruktur nutzt die am besten geeignete technische Lösung, um seine Aufgabe zu erfüllen. Das gilt auch für die Speicherung von Daten, die von den Geschäftsanforderungen und den Service-Verträgen (SLAs) abhängt. Im Gegensatz zu Data Lakes stützt sich diese neue Architektur auf ein verteiltes Konzept und richtet die Datenspeicherauswahl an der Datennutzung aus. Sie arbeitet mit mehreren Technologien, die jeweils spezifische Aufgaben erfüllen. Ein „hybrider“ Ansatz kann zudem erhebliche Einsparungen bei Kosten und Bereitstellungszeiten erzielen, wenn Änderungen oder Ergänzungen im Warehouse erforderlich sind.

Ein Begriff für diese neue Architektur ist Logical Data Warehouse, ein anderer virtueller Data Lake. In beiden Fällen gibt es keine zentrale Datenbank. Das Logical Data Warehouse ist ein Ökosystem mehrerer zweckgebundener Datenbanken, Technologien und Tools. Sie arbeiten synergetisch zusammen, um die Datenspeicherung zu verwalten und leistungsfähige Business-Analysefunktionen bereitzustellen.

Der ursprüngliche – nicht erfüllte – Anspruch an das traditionelle Data Warehouse bestand darin, Informationen mit einer einzigen Abfragesprache abzurufen, schnelle Abfrageergebnisse zu erhalten und verschiedene, für spezifische Zwecke erstellte Datenmodelle oder -ansichten rasch zusammenführen zu können. Das Logical Data Warehouse erreicht alle drei Ziele, ohne sämtliche Daten an einen zentralen Speicherort zu kopieren oder zu verschieben – durch die Kombination des Data Federation-Konzepts mit der physischen Datenintegration und einer gängigen Abfragesprache (SQL).

Die physische Datenintegration ist eine robuste Funktionalität des Logical Data Warehouse, die für schnelle Abfrageergebnisse sorgt. Gleichzeitig wird die Performance vom Quelldatenspeicher entkoppelt und richtet sich nach der Datenbank des Logical Data Warehouse. So wird der aufwendige physische Transfer der Daten auf ein Minimum reduziert und vereinfacht. Größere Verzögerungen durch Datenverschiebungen im kritischen Pfad von Datenintegrationsprojekten lassen sich damit effektiv vermeiden.




Relational database management system (RDBMS)
Hadoop Distributed File System (HDFS)
Resource Description Framework (RDF)

Source: Gartner (September 2014)

data virtuality any data. any system. any time.

In dem Bericht *„Understanding the Logical Data Warehouse: The Emerging Practice“* äußerte sich Gartner zu diesem Konzept. Gartner verweist darauf hin, dass es Unternehmen große Flexibilität bietet, die unterschiedliche Datenanforderungen zu unterschiedlichen Zeiten haben. In vielen Fällen sei beispielsweise eine zentrale Datenbank wie ein traditionelles Data Warehouse oder eine Analysedatenbank erforderlich, wo häufig benötigte Daten oder auch die Daten mit der kürzesten Abfragezeit gespeichert und zu Performance-Zwecken optimiert werden könnten.

Immer häufiger brauchen Analysten die Möglichkeit, Daten bedarfsweise und mit einer angemessenen, garantierten Geschwindigkeit abfragen zu können. Beispiele in diesem Zusammenhang sind die Sentimentanalysen oder Analysen zur Betrugserkennung. Hier ist eine verteilte Technologie wie Hadoop erforderlich, um die riesigen Datenmengen speichern zu können, die von Social Media Feeds und Clickstream-Aktivitätsprotokollen generiert werden. Sie erfordern zudem den direkten Zugriff auf Datenquellen via Data Federation. Wie Gartner zutreffend feststellt, benötigen diese Technologien eine übergeordnete logische Schicht. Diese Schicht soll zum einen die Architektur vereinheitlichen und zum anderen ermöglichen, dass Abfragen und Prozesse nach Bedarf auf allen Systemen parallel ausgeführt werden können.

Als erstes logisches Data Warehouse stellt DataVirtuality eine solche Schicht bereit. Sie vereinheitlicht die Datenspeicher und unterstützt die von Gartner genannten Anwendungsszenarien. Da die Abfragen im Hintergrund bedarfsweise an die einzelnen Datenspeicher weitergeleitet werden, bietet die Lösung von DataVirtuality große Vorteile für Business-User. Ein und dieselbe Plattform kann für viele verschiedene Szenarien genutzt werden, sehr viel mehr als beispielsweise bei einem traditionellen Data Warehouse. Auch neue Datenintegrationsansätze sind möglich, so dass sich die Nutzer primär an den Unternehmensanforderungen ausrichten und die technologische Plattform bei Bedarf anpassen können.

Durch die Entkopplung der semantisch einheitlichen Datenzugriffsschicht (über die der Nutzer interagiert) von den eigentlichen Datenquellen wird verhindert, dass Änderungen an der ursprünglichen Datenquelle die Analyseprozesse beeinträchtigen. Der endgültige Abschied von alten Datenzugriffsstrategien gestattet die bequeme und einfache Interaktion – und ermöglicht dem Nutzer sich auf sein Ziel, statt auf die technologische Basis zu konzentrieren.

DataVirtuality führt relationale und nicht relationale Datenquellen wie beispielsweise Echtzeitdaten zusammen und erlaubt die sofortige Analyse über die Abfragesprache SQL. DataVirtuality ist die Datendrehscheibe, über die ein freier Datenaustausch zwischen allen analytischen oder operationalen Datenquellen möglich ist.

Über integrierte Konnektoren können Daten umgehend durch Analyse-, Planungs- oder Statistiktools verarbeitet oder auch in das Quellsystem zurückgeschrieben werden, je nach Bedarf. Das Logical Data Warehouse passt sich zudem Änderungen in der IT-Landschaft und dem Nutzerverhalten automatisch an. Und bietet ein Maximum an Flexibilität und Geschwindigkeit bei minimalen Verwaltungskosten.

In einem Logical Data Warehouse-Projekt lassen sich mit wenigen Klicks sämtliche datengenerierenden und -verarbeitenden Systeme wie ERP- und CRM-Plattformen, Online-Shops, Social-Media-Anwendungen und nahezu jede beliebige SQL- und Nicht-SQL-Datenquelle problemlos anbinden – in Echtzeit.

Die Nutzer haben sofort Zugriff auf die Daten – und können so lange mit diesen Verbindungen und Joins experimentieren, bis sie die gewünschten Ergebnisse erzielen.

Der wesentliche Unterschied zu traditionellen ETL-Lösungen: Beim Logical Data Warehouse müssen die Daten zur Analyse nicht bewegt werden. Das spart Zeit und Kosten bei der Entwicklung und der Datenbankstrukturierung. Schnell und flexibel, folgt das Logical

Data Warehouse einem komplett anderen Datenintegrationsansatz als das unflexible traditionelle Data Warehouse.

## Funktionsweise

Das Logical Data Warehouse verbindet zwei unterschiedliche Technologien auf intelligente Weise zu einer völlig neuen Form der Datenintegration. Die erste Technologie ist die Data Federation, die zwei oder mehr voneinander getrennte Datenbanken verbindet und sie so darstellt, als wären sie eine einzige. Die zweite ist das analytische Datenbankmanagement mit einer nutzerfreundlichen Benennung der Datenelemente und einer Modellierung, die flexible Datenaufnahme- und Modellierungsoptionen bietet.

Die Ergebnisse können sich sehen lassen. Data Federation allein ist flexibel, kann aber nicht skaliert werden. Analytisches Datenbankmanagement lässt sich gut skalieren, ist aber unflexibel. Die Kombination aus beidem sorgt für einmalige Flexibilität und eine bahnbrechende Performance. Sie steht für einen vollkommen neuen Ansatz im Umgang mit Daten.

Ein Logical Data Warehouse lässt sich beispielsweise mit verschiedenen Datenquellen gleichzeitig verbinden, unter anderem mit klassischen relationalen Datenbanken wie Oracle und MS-SQL, Nicht-SQL-Datenbanken wie MongoDB oder Hadoop, spaltenorientierten Datenbanken wie Vertica oder SAP HANA sowie mit Webdiensten wie Google Analytics, AdWords, Facebook, Twitter und andere. Die resultierende integrierte Datenübersicht erscheint im Datenanalyse-Tool so, als wären die Informationen in einer zentralen SQL-Datenbank enthalten. Die Daten lassen sich anschließend mit einer gemeinsamen Sprache abfragen.

Praktisch jedes Datenanalyse-Tool, das derzeit auf dem Markt angeboten wird (Qlik, Tableau, Aqua Data usw.), kann angebunden werden. Über die virtuelle Schicht lassen sich Daten abfragen und analysieren, ohne sie aus irgendeiner Quelle laden oder kopieren zu müssen.

Diese Methode bietet eine Fülle neuer Möglichkeiten für die explorative Datenanalyse, für Data Discovery, Rapid Prototyping und intuitive Experimente. Die Nutzer erhalten in kürzester Zeit ihre Ergebnisse und können ebenso schnell ihre Datenmodelle restrukturieren. Die Erstellung gemeinsam genutzter logischer Datenansichten, wie etwa für allgemeine KPIs und Metriken, sorgt dafür, dass jeder Bericht, jede Visualisierung und jedes Abfrageergebnis den gleichen Unternehmensstandards und -definitionen entspricht. DataVirtuality ist die zentrale Datendrehscheibe, über die sämtliche Systeme und Anwendungen im Unternehmen miteinander verbunden werden. Sie ermöglicht den Datenaustausch zwischen diesen Systemen und stellt jederzeit die allerneuesten Daten bereit.

## LOGICAL DATA WAREHOUSE AUF EINEN BLICK

| + | – |
|---|---|
| • Zusammenführung von strukturierten, unstrukturierten und Echtzeitdaten dank der Kombination von Data Federation und analytischem Datenbankmanagement<br>• Daten müssen für die Analyse nicht bewegt werden<br>• Sofortverarbeitung (Analyse) oder Rückführung in Datenquellen<br>• Zentrale Datendrehscheibe verbindet alle Systeme und Anwendungen und liefert stets aktuelle Daten | • Benötigt zur vollen Effizienz mindestens 10 verschiedene Datenquellen<br>• Kein integriertes Analysetool |

data virtuality | any data. any system. any time.

# Logical Data Warehouses im Einsatz

## EIN MODERNES DATA WAREHOUSE

Für Organisationen, die Big Data und Data Warehousing intern verbinden möchten, ist das Logical Data Warehouse von großem Nutzen.

## EIN VIRTUELLES DATA MART

Mit einem logischen Data Warehouse lassen sich ohne Schwierigkeiten virtuelle Data Marts erstellen: Die primäre Dateninfrastruktur des Unternehmens wird mit zusätzlichen Datenquellen kombiniert, die für einzelne, datengesteuerte Geschäftseinheiten relevant sind. Data Mart-Initiativen lassen sich auf diese Weise sehr viel schneller umsetzen, als wenn die Daten erst in ein traditionelles Data Warehouse integriert werden müssten.

## UNTERNEHMEN IM WANDEL

Häufig ändern sich die Strukturen von Unternehmen in kürzester Zeit. Die moderne Datenintegration ermöglicht hier die schnelle Kombination der Daten verschiedener Geschäftseinheiten und bietet der Führungsebene Transparenz in puncto BI und Analyse. Diese Flexibilität ist bei strategischen Änderungen, Fusionen, Übernahmen und anderen sensiblen Operationen, die den zeitnahen Aufbau eines zentralen Data Warehouses erfordern, von entscheidender Bedeutung.

## E-COMMERCE

Die moderne Datenintegration ist eine attraktive Lösung für E-Commerce-Akteure und Einzelhändler mit vielen unterschiedlichen IT-Systemen. Ein typisches E-Commerce-Unternehmen arbeitet beispielsweise mit einem ERP-System, einem CRM-System, Webanwendungen, mobilen Apps, E-Mail-Analyseprogramme sowie Online-Marketing-, Social Media-Marketing- und anderen Tools. Mit einem Logical Data Warehouse lassen sich sämtliche Datenquellen schnell und flexibel kombinieren, um eine 360°-Sicht auf Kunden, Produkte usw. zu erhalten.

## ONLINE-MARKETING

Online-Marketing ist in hohem Maß datengetrieben und arbeitet mit Echtzeitdaten. Die einzig sinnvolle Lösung für das Management dieses komplexen Datenstroms ist ein Logical Data Warehouse: Es lässt sich problemlos mit den verschiedensten Anbietern integrieren, die Daten für das Affiliate Marketing, das Performance Marketing, die Personalisierung und andere Zwecke bereitstellen.

## DATEN IN AKTIONEN UMSETZEN

Moderne Datenintegrationsmethoden gehen einen Schritt weiter: Die Nutzer erhalten nicht nur Daten zur Analyse – sie können Informationen auch in die Datenquellen schreiben oder anhand dieser Daten konkrete Aktionen auslösen. Mit einem Logical Data Warehouse lassen sich beispielsweise Daten aus einem ERP-System, einem CRM-System und einem Online-Shop parallel analysieren, um unabhängig von den Geschäftszeiten E-Mail-Marketingkampagnen zu starten.

## ECHTZEITANALYSEN

Das Logical Data Warehouse zeichnet sich durch die hervorragende Verarbeitung von Echtzeitdaten aus. Es kann Daten flexibel modellieren und entsprechend der neuesten Analyseinitiativen ummodellieren.

## BIG DATA-INTEGRATION

Die Open-Source- und Big-Data-Lösung Hadoop eignet sich zwar für die Analyse unstrukturierter Daten und Batch-Analysen, schneidet in interaktiven Situationen aber schlecht ab. Um von einer Echtzeit-Funktionalität zu profitieren, müssen Unternehmen das traditionelle Data Warehouse mit – häufig mehreren – modernen Big Data-Tools kombinieren, wie beispielsweise ein Oracle-Warehouse mit Hadoop und Greenplum. Das Logical Data Warehouse führt diese Datenquellen in einer gemeinsamen Ansicht zusammen und liefert so eine 360°-Sicht auf Ihre Organisation – in Echtzeit.

# Fazit

Im digitalen Zeitalter ist die Nutzung großer Datenmengen das Gebot der Stunde. Nur so lassen sich intelligente Entscheidungen treffen und Geschäftsabläufe optimieren. Zwar geht unsere Fähigkeit zur Datengenerierung noch weit über unsere Kapazitäten zur effizienten Datenanalyse hinaus – doch es sind bereits große Fortschritte erzielt worden, um hier ein Gleichgewicht zu schaffen. Spannende neue Ansätze verbinden Big Data-Systeme mit traditionellen Datenstrategien: Logical Data Warehouses bieten vielversprechende Lösungen ohne die Einrichtung einer zentralen Datenbank. Dank eines Ökosystems aus mehreren Best-Fit-Datenbanken, Technologien und Tools bieten sie nun die Möglichkeit, Echtzeitdaten effektiv zu analysieren und nützliche Erkenntnisse zu sammeln. Für Unternehmen, die auf der Suche nach wertvollen Informationen Unmengen von Daten durchsieben müssen, ist dieser virtuelle „Datensee“ die ultimative Lösung: Er hilft ihnen, ihre Produkte maßzuschneidern und Kundenwünsche zu erfüllen, von denen sie bislang nichts wussten.

# Apache Spark und Hadoop

Apache Spark und Hadoop bieten gemeinsam eine beeindruckende In-Memory-Verarbeitungsleistung für Big Data-Anwendungen. Mit der In-Memory-Verarbeitung von Spark hoffte man, viele der mit Hadoop verbundenen Latenzprobleme lösen zu können. Doch beide Technologien haben ihre Grenzen und eignen sich nicht als Einheitslösung.

Apache Hadoop ist ein Open-Source-Framework, mit dem die verteilte Speicherung und Verarbeitung sehr großer Datenbestände möglich ist – Datenmengen, deren Speicherung mit den meisten anderen Speichertechnologien unwirtschaftlich wäre. Viele der früheren Einschränkungen hinsichtlich der Speicherung und Verarbeitung großer Datenvolumina fallen bei Hadoop weg, da es mit Clustern und mehreren Servern arbeitet.

Um die Daten zu verarbeiten, werden sie nicht – wie sonst üblich – über ein Netzwerk auf den Application Server verschoben. Stattdessen erfolgt die Analyse mithilfe der Hadoop-Funktion MapReduce auf den einzelnen Servern. Anschließend werden die Ergebnisse der verschiedenen Server zu einem einzigen Abfrageresultat kompiliert.

Hadoop selbst ist kein System für sich, sondern ein Ökosystem zahlreicher, miteinander verbundener Produkte, mit dem verschiedene Arten von Analysen und Operationen für jeden beliebigen Datentyp möglich sind. Als Open-Source-System wird es beständig weiterentwickelt und verbessert. Da Hadoop komplex in der Anwendung ist, entwickeln sowohl Startups als auch etablierte Unternehmen Tools, um die Arbeit mit Hadoop zu vereinfachen. Beispielsweise erforderte die Ausführung von Abfragen im Hadoop-Ökosystem ursprünglich umfassende Kenntnisse über neue und wenige bekannte Programmiersprachen wie MapReduce, Pig und Python. Dank dieser spezifischen Programmierung konnten Abfragen durchgeführt werden, die zuvor nicht möglich waren, wie etwa

Abfragen unstrukturierter Daten. Das hatte seinen Preis: Es gab weniger Programmierer, die diese Abfragen schreiben und ausführen konnten.

Mittlerweile gibt es zahlreiche Produkte, die die Analyse der in Hadoop gespeicherten Daten mit der weit verbreiteten Abfragesprache SQL erlauben.

Das klassische Hadoop-Framework ist Batch-orientiert. Es kann deshalb riesige Datenmenge relativ mühelos analysieren: Die Arbeitslast wird auf verschiedene Hadoop-Knoten aufgeteilt, die parallel arbeiten und Ergebnisse ausgeben. Doch die Analyse kleinerer Datenvolumina ist genauso komplex und erfordert ebenso viel Programmieraufwand wie große Datenmengen. Deshalb handelt es sich insgesamt um eine eher langsame Methode der Datenabfrage.

Apache Spark und verwandte Technologien versuchen, die Abfrage-Performance von Hadoop durch eine schnelle In-Memory-Datenverarbeitungsengine mit Entwicklungs-APIs zu verbessern. Ziel dieser Technologien ist es letztendlich, die Ausführung von Streaming-, Machine-Learning- oder SQL-Prozessen in Hadoop zeitnah und mit weniger benutzerdefiniertem Code zu ermöglichen.

Mit Hadoop kann fast jede beliebige Analyse durchgeführt werden – auch die Analyse enormer Datenmengen wie Sentimentanalysen oder Analysen zur Betrugserkennung. Insgesamt ist Hadoop aber nach wie vor eine eher unreife Technologie. Ihr Ökosystem ist noch nicht voll integriert und benötigt an mehreren Stellen eine spezifische Programmierung, um voll funktionsfähig zu sein. Aufgrund des hohen technischen Anspruchs und der schwierigen Anwendung wird Hadoop oft am erfolgreichsten als günstiges Datenarchiv eingesetzt.

# Über DataVirtuality

DataVirtuality GmbH entwickelt und vertreibt die Software DataVirtuality, die Unternehmen eine besonders einfache Art der Integration und Anbindung vieler verschiedener Daten und Anwendungen ermöglicht. DataVirtuality revolutioniert das technologische Konzept der Datenvirtualisierung und baut in nur wenigen Tagen ein Data Warehouse auf, das aus relationalen und nicht relationalen Daten besteht. Mithilfe integrierter Konnektoren können Daten je nach Bedarf umgehend durch Analyse-, Planungs- oder Statistiktools verarbeitet oder in die Quellsysteme zurückgeschrieben werden. Das Data Warehouse von DataVirtuality passt sich zudem Änderungen der IT-Landschaft und des Nutzerverhaltens automatisch an. Es bietet Kunden ein Maximum an Flexibilität und Geschwindigkeit – bei minimalen Verwaltungskosten. DataVirtuality entstand 2012 aus einem Forschungsprojekt des Instituts für Informatik an der Universität Leipzig und ist heute ein global agierendes Unternehmen mit Niederlassungen in Leipzig, Frankfurt und San Francisco. DataVirtuality wird finanziert durch den Technologiegründerfonds Sachsen (TGFS) und den High-Tech Gründerfonds (HTGF).

**KONTAKT:**

Dr. Nick Golovin
Firmengründer und CEO
DataVirtuality GmbH
Telefon: +49 341 2636 2258
E-Mail: **nick.golovin@datavirtuality.com**

data virtuality | any data. any system. any time.